Br. J. clin. Pharmac. (1982), **14**, 473–482

PATIENT SUBSETS AND VARIATION IN THERAPEUTIC EFFICACY

RICHARD SIMON

Biometric Research Branch, Division of Cancer Treatment, National Cancer Institute,
Bethesda, Maryland 20205, USA

1 The determination of what treatment is best for what kinds of patients is a general objective of clinical research. We consider here the extent to which this objective can be accomplished reliably in a single clinical trial.

2 The reliability of subset analyses is often poor due to problems of multiplicity and limitations in numbers of patients studied. Implications for the design of clinical trials are presented.

3 Statistical approaches to subset analysis are reviewed in a general manner. In order to obtain the degree of reliability usually demanded of clinical therapeutic evaluations, 'statistically significant' interactions between relative treatment efficacy and subsets should be demonstrated. Exploratory analyses of subset differences are important but should be reported as hypotheses to be tested in separate studies.

Introduction

Ideally, medical trials should yield reliable and precise predictions of clinical outcomes as a function of treatment and patient characteristics. This would allow appropriate choice of treatments for individual patients. Such an objective is quite ambitious however, and this article will discuss to what extent it is useful as a basis for designing randomized therapeutic clinical trials.

The strong advantages of randomization will not be reviewed here. Even among randomized studies however, one still reads conflicting reports of trials purporting to evaluate the relative benefits of the same treatments. The reasons for such inconsistencies include variability in patient selection, treatment administration, outcome evaluation, inappropriate data analysis and the statistical properties of large numbers of studies with modest numbers of patients. One major factor is the conduct of subset analyses in an effort to determine which treatment is preferable for what kinds of patients. Analysing patient subsets is a natural part of the process of improving therapeutic knowledge through clinical trials. But the naive interpretation of the results of such examinations is a cause of great confusion in the therapeutic literature. The adopted study designs and methods of statistical analysis influence the extent to which erroneous conclusions are likely to result from the use or non-use of subsets analysis. These issues will be the topic of this paper.

Patient selection

Patient selection considerations are important for two major reasons. First, they determine the target population of patients to which the major conclusions of the study will apply. Second, they strongly influence the reliability and precision of our resulting therapeutic predictions for individual members of the target population. These two aspects will be discussed in turn.

If the conclusions of a clinical trial were not generalizable to patients other than those actually included in the study, clinical research would be futile. The only statistical basis for generalization, however, is the assumption that the studied patients constitute a random sample from a larger population. We rarely, if ever, actually randomly select patients, but it generally seems reasonable to act as if we had. The patient selection criteria determine the target population, and extrapolating our conclusions beyond this has little justification. Some investigators favour broad patient selection so that the conclusions are applicable to the greatest number of patients. As we shall see, however, this approach has serious limitations.

In many clinical trials the patients are heterogeneous and the responses are variable. It is for this reason that statistics plays a major role in the conduct of these studies. Statistical analysis is no panacea, however. As we shall see later, the employment of numerous subset analyses may be accompanied by

0306–5251/82/100473–10 $01.00

increasing uncertainty about the reliability of conclusions. Unless the study was designed large enough to support the reliable separate analysis of several initially defined subsets, the basic conclusion of the study should be the overall comparison of treatments for all patients randomized. Though this overall comparison may be balanced or adjusted for prognostic factors, the conclusion is generalizable only to a mixture of the same kinds of patients studied in the trial. The overall comparison addresses the question: If I must use treatment A or B for all patients in the target population, which should I use? Though the target population may be of impressive size because of broad patient selection criteria, the conclusion applies 'on the average'.

Correct conclusions for populations can be incorrect for individuals within the population. For example, an aggressive therapy may be beneficial to young ambulatory patients but harmful to older debilitated patients. If a mixed population is studied with inadequate numbers for reliable subset analyses, the conclusion of no average positive or negative effect of therapy will be erroneous for all patients. Unfortunately, it is generally impossible to acquire the number of patients required for reliable evaluation of the many subsets that could be identified as being of interest. With numerous subsets, the required number of patients would be huge. Long before that number could be recruited, one would be pressed to terminate the study because the average relative efficacies of the treatments would be established only too well.

No two study patients are exactly alike and no future patient will be exactly alike one of our study patients. So at some point it is necessary to settle upon a target population for whom we are willing to attempt to reach a reliable overall conclusion about therapeutic effects. We may hope to perform subset analyses but should recognize that these will generally be less reliable. There will be a level of refining our patient selection criteria beyond which it is not feasible to obtain sufficient patients for a reliable study. Some refinements may reduce variability in outcome and thereby improve the sensitivity of the resulting study. Other refinements serve to eliminate patients who are likely only to dilute the sensitivity of our study for the majority of the target population. For example, to include in cancer chemotherapy studies subsets of extremely debilitated patients whose treatment is already planned to be less than others is likely to have a diluting effect. Such subsets are often included because some statistical table said that *n* patients were needed and it didn't specify what kind.

The argument for relatively narrow patient selection criteria was emphasized by Sir A.B. Hill in his description of a Medical Research Council trial of streptomycin (Hill, 1951): 'In short, the questions asked of the trial were deliberately limited and these closely defined features were considered indispensable, for it was realized that no two patients have an identical form of the disease and it was desired to eliminate as many of the obvious variations as possible. This planning . . . is a fundamental feature of the successful trial. To start out upon a trial with all and sundry included, and with the hope that the results can be sorted out statistically in the end is to court disaster.'

Some statisticians today do not agree with this viewpoint (Peto *et al.*, 1976). They say, essentially, do not waste time arguing about whether subsets of patients should or should not be included in the trial; if it seems reasonable to include them, then do so. For extremely large trials where reliable subset analyses are possible, this viewpoint is reasonable. In other cases it provides fuel for questionable conclusions based upon inadequate numbers, has a diluting effect on the primary comparison and adds subjectivity to the analysis.

Stratification

Even when the above recommendations are followed, there will generally remain some heterogeneity in the target population with regard to known prognostic factors. Stratified randomization is often used to ensure a greater degree of balance of the treatment groups with regard to these known prognostic factors than can be ensured by pure randomization.

Stratified randomization is usually accomplished by partitioning the patients into mutually exclusive subsets based upon pre-treatment characteristics thought to affect prognosis. Within each subset, or stratum, a pseudorandomization is performed. One kind of pseudorandomization is the permuted block design. A permuted block of length 6 for two treatments, A and B, is a sequence of three As and three Bs. In general for two treatments, a permuted block of length 2k is a sequence of As and Bs containing exactly kAs and kBs. The treatment assignment is determined by a succession of randomly selected permuted blocks within each stratum. If there are three treatments, permuted blocks of k instances of each of three letters are used. Generally, the permuted blocks of distinct strata are prepared independently of each other.

Though limited stratification is generally desirable, over-stratification can be detrimental to a trial. With numerous strata, many will not contain enough patients by the conclusion of the trial to complete permuted blocks. Consequently, balance with regard to the most important factors may be impaired by the inclusion of secondary factors. Overstratification in the extreme becomes equivalent to no stratification at all (Simon, 1980). Details about stratification and review of new methods that accommodate more stratification variables are described by White & Freedman (1978), Pocock (1979) and Simon (1979). The purpose of stratification is to ensure reasonable balance between the treatment groups with regard to prognostic factors. Stratification does not imply that separate evaluation of relative treatment efficacy will be performed within each stratum.

It is often desirable to incorporate the stratification factors into the analysis, in order to increase the sensitivity of the trial. Treating the known sources of variability as unknown sources of noise is to be avoided when possible. To help clarify this point, consider the illustration in Figure 1. It is assumed that there are two strata and two treatments. The upper graph depicts the frequency function, or histogram, of survival for each of the two treatment groups within each of the two stratum. For purposes of illustration, the concept of censored survivals is ignored. It is clear that stratum II patients have a better prognosis than stratum I patients, because their frequency functions are displaced to the right. Also, within each stratum treatment B seems more effective than treatment A. The desirable analysis consists of calculating an estimate Δ_I of the treatment difference within stratum I and an estimate Δ_{II} of the treatment difference within stratum II. A weighted average of Δ_I and Δ_{II} is used as the overall test statistic for evaluating the statistical significance of the treatment difference. Of course significance tests could be performed within each stratum, but the limited number of patients usually makes this undesirable because the power of the test will be small. The variability associated with the weighted average of Δ_I and Δ_{II} is estimated based upon the variability for each of four curves at the top of Figure 1. The treatment differences in the case shown are reasonably large relative to the variability of the individual curves, so if the number of patients is not too small, the comparison will be found statistically significant.

The bottom illustration is intended to represent a redrawing of the top illustration, except that the stratum distinctions are ignored. Each of the two curves represents a frequency function of survival for a pooled treatment group consisting of patients from both strata. Though the two treatment groups may be perfectly 'comparable' with regard to the stratification variable, if the pooled groups are compared directly a significance test having poor power will result. Thus, when important prognostic variables can be incorporated into the analysis, the random fluctuations caused by such variables are in a sense eliminated from comparison of treatments. Major improvements in sensitivity are only possible for strong prognostic variables. However, it is still the average relative benefit of the treatments for our target population being evaluated. Some prominent statisticians believe that, except for small studies, stratification is an unnecessary complication (Peto *et al.*, 1976). They point out that stratified analysis can be performed regardless of whether the randomiza-

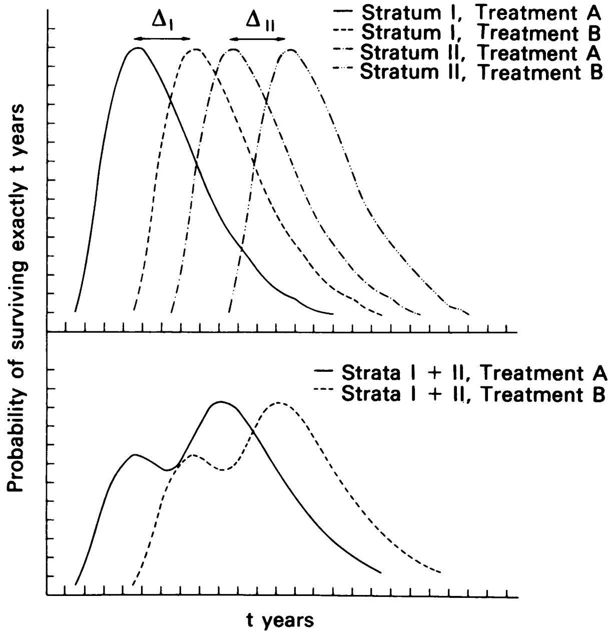

Figure 1 Use of a prognostic factor in analysis.

tion procedure was stratified and that stratified randomization contributes little to the average power of the resulting statistical significance tests. It must be remembered however that stratification does protect against chance imbalances that may severely impair the basic credibility of the study and the ability to perform a stratified analysis (Lasagna, 1955; Brown, 1980). The fact that these chance events are of low probability will be of little comfort if they materialize in a study.

One point that should be emphasized is that availability of stratification methods is no justification for broadening the patient selection criteria. Even where it is possible to balance the groups with regard to prognostic heterogeneity, this is generally accomplished at a loss of statistical precision and a loss of specificity concerning the relevance of conclusions to individuals within the target population. Stated differently, the constant relative treatment efficacy shown in Figure 1 is not generally the case.

Interpretation of subset analysis

The basic problems of interpreting separate analyses of relative treatment efficacy within patient subsets can be illustrated as follows. Suppose that we have randomly assigned treatments A and B and partition our patients into G mutually exclusive subsets. For each subset we perform a statistical significance test and declare a difference 'statistically significant' if the calculated significance level is α or smaller. We obtain one 'significant' difference using $\alpha = 0.05$, but someone points out that even if the treatments are identical the probability of obtaining at least one significant result is

$$1 - (1 - \alpha)^G. \quad (1)$$

For $\alpha = 0.05$ and $G = 10$ this probability is about 0.40. That is if the treatments are equivalent for all ten subsets, there is a 40% chance that at least one difference will appear 'significant' at the 0.05 level. For five subsets the probability is about 23%.

For many clinical trials, the number of subsets which might be examined is far greater than ten. Multiplying the large number of possible subsets with the number of outcome measures and the number of treatment pairs (for studies including more than two treatments) can result in a very large number of possible comparisons. These comparisons will generally not be independent and expression (1) becomes an upper bound rather than an exact formula for the probability of obtaining at least one 'significant' result by chance alone. Nevertheless, with many comparisons this probability approaches unity. For G independent comparisons, the expected number of 'significant' differences by chance alone is $G\alpha$. The crux of the problem of performing many comparisons is thus that it may not be reasonable to interpret individual results at face value.

The most frequent approach for dealing with this problem is the following. Those comparisons of greatest *a priori* interest are specified at the outset of the study. These should be few enough in number so that a troublesome number of comparisons is not produced; alternatively, a statistical method is adopted which adjusts the individual significance levels for the number of comparisons. For example, the protocol might state that the major comparisons are those of the entire two treatment groups with regard to survival, complete response rate, duration of responses and toxicities. If the resulting data suggests treatment differences for other comparisons, for example only within a few subsets, this would be reported as suggestive evidence of a lower order of reliability to be evaluated in a subsequent trial. Thus, the data analysis is partitioned to yield two dividends: the testing of a few pre-specified hypotheses, and the generation of new hypotheses by subset analyses. Clinicians often fail to make this distinction adequately in reporting work. For example, Lasagna (1976) comments: 'I wish the authors had commented further on the practice of parcelling out subgroups of patients who allegedly are specifically benefited (or hurt) by treatment, despite the absence of statistical differences between the total groups. This has occurred in both the UGDP and Coronary Drug Projects. The probability calculations are by no means clear in such cases, and one would have thought that the statisticians of the world would have rallied round the flag and decried such tactics as anything other than a source of hypotheses to be tested prospectively in new controlled trials. What we have seen, instead, is the use of such subanalyses to come to conclusions in which regulatory decisions are proposed.'

Introductory statistical courses are generally oriented to the Neyman-Pearson theory of testing pre-specified hypotheses. In this theory a statistical significance test is a decision rule for accepting or rejecting hypotheses. Many applied statisticians see the role of analysis more broadly as summarizing data, estimating effects, and quantitating weight of evidence (Fisher, 1955; Anscombe, 1963; Cutler *et al.*, 1966). A significance level of 0.07 conveys more information than merely that the null hypothesis should be accepted because the pre-specified type 1 error was 0.05. The distinction between one-sided and two-sided significance levels becomes critical in the Neyman-Pearson framework, because it may represent the difference between accepting and rejecting the null hypothesis. For purposes of summarizing data and weight of evidence, the distinction is not critical (though one should always state which type is used) because the 0.05 cutoff value has no unique meaningfulness. The hypothesis testing approach is also not really an adequate framework for medical decision making, because it forces decisions on the user where there is insufficient evidence. For comparing treatment efficacy it does not consider whether differences are of practical importance, the costs and complications of each therapy, the losses consequent on wrong decisions or prior knowledge (Cox, 1958).

Although the Neyman-Pearson theory is not really adequate for describing weight of evidence or for medical decision making, the concept of dichotomizing hypothesis testing verus hypothesis generation has several benefits. First, the major comparisons can be addressed without concern that the findings are the result of ransacking the data for suggestive findings. Second, it focuses attention at the design stage of the deleterious effects of starting 'out upon a trial with all and sundry included . . . ' as stated by Hill. And finally if results are intelligently and cautiously reported, it frees one to examine subset results in the data.

Basic methods for subset analysis

It is widely accepted that claims of therapeutic differences for subsets of patients are unreliable unless either an overall difference in therapeutic benefit for all patients studied has been demonstrated or strong quantitative evidence is presented that therapeutic benefit varies among *a priori* defined subsets. We first examine the latter condition.

Demonstrating treatment-subset interactions

Suppose that two treatments are compared in a study and that there are G mutually exclusive subsets of patients which we are interested in examining. For the method to be described, it is important that the subsets not be selected as a result of perusing the data. They should represent subsets with reasonable numbers of patients of *a priori* interest. The usual analysis of variance model is

$$X_{igt} = a_g + b_t + c_{gt} + e_{igt}. \qquad (2)$$

X_{igt} represents the observed outcome for the i'th patient in subset g who receives treatment t. $a_1, a_2, \ldots, a_G$ are unknown constants which represent the prognostic effect of being in subset g. b_1 and b_2 are unknown constants that represent the average influences of treatments upon response. The unknown constants c_{g1} and c_{g2} represent modifications to average treatment effects for patients in subset g. The e_{igt} are assumed to be independent normally distributed random variables wth mean zero. They represent experimental variability unrelated to treatment or subset effects.

The c_{gt} terms in equation (2) are called interactions. If they are all zero then the influence of treatment upon survival is uniform for all subsets. In the usual analysis of variance, the first hypothesis tested is that $c_{g1} = c_{g2} = 0$ for all subsets g. Armitage (1971) describes how this no interaction hypothesis can be tested. If the data provides strong evidence for rejecting this hypothesis then one has a reasonable basis for abandoning model (2) and comparing treatments separately by subset. Generally, one would attempt to identify mutually exclusive classes of subsets within which there are no interactions. The subsets within a class would be pooled, but the classes would be analysed separately. If there is not strong evidence for rejecting the no-interaction hypothesis, then subset analyses are not justified. Model (2), with the c terms omitted, would be the basis for analysis. In this analysis the prognostic influences of subsets are included, but it is only the average treatment effects for all patients that are compared.

The above approach to subset analysis is useful even when the specific model (2) is not appropriate. For many clinical trials the major outcomes are either binary (e.g. complete remission or no complete remission) or incompletely observable (e.g. survival). Survival data is called partially 'censored'. For a patient alive at the time of analysis, say four years after entry to the study, we know only that his or her survival is at least four years. The true survival is 'censored' at four years.

For the binary response case, the analogue of (2) usually adopted is the logistic model

$$\log(p_{igt}/q_{igt}) = a_g + b_t + c_{gt}. \qquad (3)$$

p_{igt} represents thę probability of a successful outcome for the i'th patient in subset g who receives treatment t, and $q_{igt} = 1 - p_{igt}$. The other unknown constants are as before. The no-interaction hypothesis that all c equal zero can be tested using methods described by Cox (1970). As before, strong evidence against this hypothesis is a basis for comparing treatments separately within subsets.

In the remainder of this paper several approximate methods of analysis are presented. This information is not an adequate substitute for collaboration with a biostatistician. The statistical problems addressed in this paper are non-trivial, and experimentation with human subjects is incompatible with using inferior methods of analysis. Use of the approximate methods under the conditions described, however, is better than the frequent practice of totally ignoring the issues discussed here.

With large sample sizes, an approximate interaction test can be performed in the following way for the logistic model. Let S_{g1} and F_{g1} denote the number of successes and failures respectively for patients in subset g receiving treatment 1. Let S_{g2} and F_{g2} denote these numbers for patients receiving treatment 2.

Add 1/2 to each of these numbers and call the results s_{g1}, f_{g1}, s_{g2} and f_{g2}. Let

$$\hat{\psi}_g = \log (s_{g1}f_{g2}/f_{g1}s_{g2})$$
$$w_g = 1/(1/s_{g1} + 1/f_{g1} + 1/s_{g2} + 1/f_{g2}) \qquad (4)$$

and

$$\bar{\psi} = \Sigma\, w_g\, \hat{\psi}_g / \Sigma\, w_g \qquad (5)$$

where the summation is over the subsets g and natural logarithms are used. For a patient in subset g the ratio of probability of success to probability of failure with treatment t, p_{gt}/q_{gt}, is called the odds for success. Dividing the odds for treatment one by the odds for treatment two gives the odds ratio for that subset. The quantity $\hat{\psi}_g$ above estimates the log odds ratio for subset g. The quantity $1/w_g$ is the approximate variance of $\hat{\psi}_g$. The quantity $\bar{\psi}$ is an average log odds ratio over all the subsets. If all the odds ratios are equal, then there is no interaction and the relative treatment effect is the same for each subset. This can be approximately tested by calculating

$$Z = \Sigma\, w_g(\hat{\psi}_g - \bar{\psi})^2 \qquad (6)$$

where the summation is over all G subsets. If there is no interaction then Z should approximately have a chi-square distribution with G−1 degrees of freedom. For example with 10 subsets, if Z exceeds 16.9 then the approximate significance level is less than 0.05 and one has a reasonable basis for examining how relative treatment efficacy varies among subsets. This test is only adequate for relatively large sample sizes. Unless each treatment group within each subset contains at least five successes and five failures, this approximation should not be used. Even when these minimal requirements are met, the approximation is not considered entirely adequate (Fleiss, 1979) and the more complicated methods of Cox (1970) should be employed for definitive analysis. The method presented is useful, however, for exploring whether more accurate analysis is warranted.

For survival data (or other failure time data) a very commonly used model is the proportional hazards model of Cox (1972). For the subset problem this model takes the following form:

$$\lambda_{gt}(\tau) = f(\tau)\exp(a_g + b_t + c_{gt}). \qquad (7)$$

The function $\lambda_{gt}(\tau)$ represents the force of mortality at time τ for a patient in subset g who received treatment t. It can be thought of as the failure rate or the probability of death at time τ for such a patient alive just before τ. This function is called a hazard function of a survival distribution. In this model, the hazard function equals some unknown function of time $f(\tau)$ times an exponential function that depends upon the unknown constants introduced before. As for the previous models, one can test the no-interaction hypothesis that all c equal zero. If the treatment effects are not large, the method of Haybittle (1979) can be used to perform an approximate evaluation of the presence of interaction. One calculates

$$\hat{\psi}_g = (O_{g1} - E_{g1})/V_{g1}$$

and $\qquad (8)$

$$w_g = V_{g1}$$

where O_{g1} is the observed number of failures with treatment 1 in subset g, E_{g1},V_{g1} are the Mantel-Haenszel (Mantel, 1966) expectation and variance of the number of failures within treatment 1 in subset g under the null hypothesis that the treatments are equivalent. Using expressions (5), (6) and (8), an approximate test for the absence of interactions can be performed for the proportional hazards model. The expression for $\hat{\psi}_g$ given in (8) is an estimator of the logarithm of the ratio of hazard functions (failure rates) for treatment 1 relative to treatment 2 within subset g.

Unfortunately the concept of interaction is model dependent. For example, with binary responses the difference in success probabilities between treatments $p_{g1}-p_{g2}$ may be constant for all subsets. This would represent an interaction in the odds ratio model. In practice this is generally not a serious problem. The power of the interaction tests will often be limited to detecting major reversals of treatment preference, or a strong treatment preference in one subset in the presence of treatment equivalence for the others. *A priori*, the presence of important true interactions is unknown whereas the spurious appearance of variability in relative efficacy among many subsets is almost a certainty. Consequently it is prudent to require quantitative documentation that such apparent differences are not the result of seeking out chance fluctuations. When 'statistically significant' interactions are found, however, one must examine whether it is just a matter of scale of measurement upon which the model is based.

Tests of average treatment effects

If both an average benefit of one treatment for all study patients and a significant interaction can be demonstrated, then careful examination of the variation in relative benefit among subsets is warranted. Some statisticians make the demonstration of an average benefit a strong requirement. For example (Peto *et al.*, 1976): 'The fundamental *P*-value to be reported is the overall comparison of treatments adjusted for retrospective stratification. If this is not significant, it is unwise to conclude without expert statistical assistance that any treatment differences in individual strata are real.'

If subsets are only examined when the overall *P* value is less than 0.05, then in 95% of the trials for which there are no real treatment differences, no apparently 'significant' differences will be claimed, even for subsets. Nevertheless, it seems prudent not to infer strong conclusions about differences in relative benefit among subsets unless a significant interaction can be demonstrated, regardless of whether there is an average benefit of one treatment.

The frequently used Mantel-Hanszel test (Mantel, 1966) for an overall comparison of survival adjusted for the prognostic effect of subsets is equivalent to calculating

$$\bar{\psi}^2 \Sigma w_g \qquad (9)$$

based on expressions (5) and (8). Under the null hypothesis of treatment equivalence within each subset, the above quantity approximately has a chi-square distribution with one degree of freedon. The same method can be used for binary responses.

Examining subsets

The most commonly used approach to examining how relative benefit varies among subsets is use of separate hypothesis tests for treatment differences within subsets. This approach may give very misleading results. The statistical significance level is a function of the sample size as well as the observed difference. Hence a 'non-significant' result for one subset and a 'significant' result for another subset may correspond to exactly the same observed treatment differences. Unless the evaluation of treatments within a subset was initially identified as a major goal of the study, it is unlikely that sufficient numbers of patients will be available for reliable comparisons of this type. Use of the 'significant' or 'non-significant' dichotomy of the Neyman-Pearson theory within subsets, will often yield firm decisions where reliable conclusions are not possible. It is also possible to spuriously eliminate a strong overall treatment effect by decomposing the data into many subsets for separate analysis.

Calculations of confidence intervals for the treatment differences within subsets is likely to be less misleading than hypothesis testing. Confidence intervals exhibit a range of true treatment differences consistent with the subset data and offer less temptation to confuse 'not significant' with 'not different'. For the binary response model (3), an approximate 95% confidence interval for the log odds ratio within subset g is $\hat{\psi}_g \pm 1.96/\sqrt{w_g}$ given by (4). A log odds ratio of zero corresponds to treatment equivalence. The same expression based on (8) provides an approximate confidence interval for the logarithm of the ratio of failure rates for temporal data.

Since ψ_g represents in general a measure of relative treatment efficacy for subset g, the statement that relative efficacy for subset g_1 differs for that of subset g_2 is equivalent to the statement $\psi_{g1} \ddagger \psi_{g2}$. For any class consisting of G′ subsets (2 or more), we can test the hypothesis of uniformity of relative efficacy among this class by testing the hypothesis that the ψ_g within the class are equal. This can be done in the following way. Define ψ as in equation (5) but with the summation over the class in question (perhaps over two subsets). Z is calculated from (6) with the same limitation of the summation. Under the hypothesis of constancy of relative efficacy among the G′ subsets in the class, Z will approximately have a chi-square distribution with $G' - 1$ degrees of freedom.

With this approach one can attempt to identify classes of subsets for which relative efficacy cannot be demonstrated to differ. For such a class C, the average relative treatment efficacy, $\bar{\psi}_C$, is given by equation (5) with the summations over the subsets within C. The approximate standard error of $\bar{\psi}_C$ is

$$\sigma_C = (1/\Sigma\, w_g)^{1/2}$$

again with the summation over C. One can thus test whether the true pooled $\psi_C = 0$ and calculate approximate 95% confidence limits for ψ_C as $\bar{\psi}_C \pm 1.96\ \sigma_C$. More accurate methods are described by Fleiss (1979) and Cox (1970, 1972). The data will often not be definitive enough to permit mutually exclusive classes C to be uniquely defined. The interaction tests will generally suffer the same lack of power as the comparison of treatments within individual subsets. Nevertheless, if we cannot demonstrate a difference in relative efficacy between two subsets, it is more reasonable to utilize their data in a reinforcing way than to force both within subset treatment comparisons to suffer from lack of power.

The same type of analysis can be peformed with survival data without adopting the proportional hazards model. Suppose, for example, we adopt as one measure of outcome the probability of surviving at least two years from start of treatment. We can estimate the probability p_{gt} for patients in subset g receving treatment t using the Kaplan-Meier method as described in Peto *et al.* (1976). Although this estimate will generally utilize survivals 'censored' before two years, Peto *et al.* (1976) also point out that the estimate is about as precise as if we had r_{gt}/p_{gt} total patients all followed for a complete two year period where r_{gt} is the number of patients actually alive and followed for at least two years. Consequently the analysis of variation of treatment influence on two-year survival can be approximately performed by the binary response methods outlined above. We assume that there are r_{gt} 'successes' and $(1 - p_{gt})r_{gt}/p_{gt}$ 'failures' among the patients in subset g receiving treatment t.

If there are numerous subsets, then the general approach outlined above may produce spurious indications of differences in relative efficacy. For example suppose that there are 50 subsets, that relative treatment efficacy is the same for 49 but very different for one and we obtain a statistically significant overall interaction test. There will be 1176 (= 49 × 48/2) possible pairwise comparisons among subsets in which relative efficacy is actually the same. If these comparisons were independent, we would expect about 59 spurious claims of 'statistical significance' at the 5% level. Since the comparisons are dependent, the expected number may be less than 59, particularly since some subsets may be small. But clearly, the 5% criterion is inadequate. A variety of 'multiple comparison' procedures have been developed to deal with this problem (Miller, 1966). If there are no more than 10 subsets, however, and if pairwise comparisons are only performed when the overall interaction test is significant at the 0.05 level, then the general approach outlined above (though not necessarily the particular methods) should be adequate.

For those who insist on performing hypothesis tests within each subset when no average treatment benefit or no overall interaction is demonstrated, and for those of us who must continue reading their papers, a reduced significance level is required to avoid numerous false positive results. In order to ensure a probability no greater than 0.05 that at least one of G treatment-within-subset comparisons will be declared significant by chance alone, it is sufficient that a cutoff level of 'significance' of 0.05/G be used for each individual comparison (Tukey, 1977).

Determining subsets

The discussion above assumes that certain natural subsets are of *a priori* interest. In attempting to better understand therapeutic results, however, it is reasonable to search a large body of data for ways to define subsets for which one treatment or the other is superior. Though no theory of how to best do this has been devised, Byar & Corle (1977) have developed and illustrated the use of multivariate regression methods for this purpose. There is subjectivity in this process and methods of estimating reliability of the predictions have not been adequately developed. Byar & Corle (1977) comment: 'Possibly the greatest value of this sort of analysis is heuristic—it suggests relationships which might not have been apparent by more conventional methods. The proof of any conclusions tentatively drawn must depend on future experiments designed specifically to test the results suggested by the analysis.'

The thorough examination of data resulting from a good clinical trial is certainly warranted. The development of new statistical tools for the exploratory analysis of subsets is a useful area of research. The good advice of Byar & Corle (1977) should be remembered, however, in reporting or reading the results of such an analysis.

Subsets determined after the start of treatment

The previous sections have assumed that the patient subsets are determined by baseline characteristics of the patient and his or her disease. In some studies subsets are determined by events occurring after the start of therapy. For example, Bonadonna & Valagussa (1981) have attempted to evaluate adjuvant chemotherapy for the subset of postmenopausal breast cancer patients who received adequate doses. Other reports attempt to analyse the subset of patients who fully comply with treatment administration. A danger with such analyses is that the subset definition may inadvertently select non-comparable patients for the two treatments. One of the largest differences in mortality reported by the Coronary Drug Project was between the good compliers and poor compliers in the placebo group. This striking finding could not be adequately accounted for by imbalances in about 40 baseline variables considered (Coronary Drug Project, 1980). Consequently one should be cognizant of possibility for bias in the evaluation of such subsets (Canner, 1981).

Conclusion

An important objective for clinical research is to determine reliably which treatment is best for what kinds of patients. This broad objective is not achievable in most clinical trials, however, because of limitations in the number of patients studied. More realistic goals for most clinical trials are: (1) develop reliable conclusions about average relative treatment efficacy for groups of patients selected initially on coherent clinical grounds; and (2) generate hypotheses to be tested in later studies about relative efficacy for subsets of patients. If the study is planned so that adequate numbers of patients are recruited within several subsets defined *a priori*, or if there are obvious differences in relative effiacy among such subsets, then greater progress towards the broader objective is possible.

The hypothesis generation part of the analysis is very important, but findings are often reported as being definitive. The problem of subset multiplicity is often not acknowledged, and authors seldom try to measure the strength of evidence for variations in relative treatment efficacy among subsets. It is not dishonorable to thoroughly examine carefully collected data for therapeutic leads. Statistical naiveté and failure to exercise caution and sound scientific skepticism in reporting results, however, adds confusion to the therapeutic literature.

The difficulties of deriving reliable conclusions about therapeutic efficacy for subsets should be recognized in developing the patient selection criteria for the study. One should try to ensure that a clear picture of the average therapeutic benefit for a medically meaningful target population taken as a whole will be obtained or that adequate numbers of patients are accrued for independent analysis of pre-specified subsets. Too many studies include very heterogeneous subsets of patients because the investigators want to do a study that is really not feasible with the small number of patients they see. Conclusions from subset analyses generally are of a lower order of reliability than the conclusion of average benefit. The latter must not be sacrificed in an attempt to achieve too much.

Tukey (1977) distinguishes a 'clinical inquiry' from a 'focused clinical trial'. In the former we study heterogeneous patient populations and hope to determine the preferred treatment for many subsets. In the latter, only a single type of outcome is to be evaluated for a single defined population of patients. Because of problems inherent in interpreting multiple subset analyses, he concludes: '. . . I do not believe that a clinical inquiry, by itself, is likely to be an ethically satisfactory means of providing definitive evidence that an intervention or therapy is an improvement . . . It is right for each physician to *want* to know about the behavior to be expected from the intervention or therapy when applied to his particular individual patient . . . It is not right, however, for a physician to *expect* to know this—except, possibly, for the most dramatically effective and time-tested interventions or therapies . . . Knowing that, for one class of patient, a clinical inquiry has reached some specific level of significance, such as 4%, is not evidence of the same strength as knowing that a focused clinical trial, involving a single prechosen question, has reached exactly that level of significance . . .'

Few clinical studies are pure 'focused clinical trials' as described by Tukey. It is possible to perform 'clinical inquiry' while achieving reliable conclusions for the few major questions addressed. Broadening the focus too much, however, is achieved at the expense of decreased reliability for all findings. The primary questions and the primary target populations should be distinguished in the protocol and one should ensure that sufficient numbers of patients of identified types are studied. Other analyses and patients can then be included and results thoroughly examined as clinical inquiry.

References

ANSCOMBE, F.J. (1963). Sequential medical trials. *J. Am. Statist. Ass.*, **58**, 365–383.

ARMITAGE, P. (1971). *Statistical Methods In Medical Research*. Oxford: Blackwell.

BONADONNA, G. & VALAGUSSA, P. (1981). Dose-response effects of adjuvant chemotherapy in breast cancer. *New Engl. J. Med.*, **304**, 10–15.

BROWN, B.W. Jr. (1980). Statistical controversies in the design of clinical trials. *Controlled Clin. Trials*, **1**, 13–27.

BYAR, D.P. & CORLE, D.K. (1977). Selecting optimal treatment in clinical trials using covariate information. *J. chronic Dis.*, **30**, 445–459.

CANNER, P.L. (1981). Influence of treatment adherence in the Coronary Drug Project (letter). *New Engl. J. Med.*, **304**, 612–613.

CORONARY DRUG PROJECT (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *New. Engl. J. Med.*, **300**, 1038–1041.

COX, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, **29**, 357–372.

COX, D.R. (1970). *Analysis of Binary Data*. London: Methuen.

COX, D.R. (1972). Regression models and life tables. *J. Roy Statist. Soc. B*, **34**, 187–220.

CUTLER, S.J., GREENHOUSE, S.W., CORNFIELD, J., SCHNEIDERMAN, M.A., ZELEN, M., SHAW, L.W. & BEEBE, G.W. (1966). The role of hypothesis testing in clinical trials. *J. chronic Dis.*, **19**, 857–882.

FISHER, R.A (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc. B*, **17**, 69–78.

FLEISS, J.L. (1979). Confidence intervals for the odds ratio in case-control studies: The state of the art. *J. chronic Dis.*, **32**, 69–77.

FLEISS, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: Wiley.

GART, J.J. (1971). The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Review Inter. Stat. Inst.*, **39**, 16–37.

HAYBITTLE, J.I. (1979). The reporting of non-significant results in clinical trials. In *Clinical Trials in 'Early' Breast Cancer*. eds Scheurlen, B.R., Weckesser, G. & Armbruster, I., pp. 28–39. Berlin: Springer-Verlag.

HILL, A.B. (1951). The clinical trial. *Br. med. Bull.*, **7**, 278–282.

LASAGNA, L. (1955). The controlled clinical trial: Theory and practice. *J. chronic Dis.*, **1**, 353–367.

LASAGNA, L. (1976). Randomized clinical trials (letter). *New Engl. J. Med.*, **295**, 1086–1087.

MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719–748,

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, **50**, 163–170.

MILLER, R.G. Jr. (1966). *Simultaneous Statistical Inference*. New York: McGraw-Hill.

PETO, R., PIKE, M.C., ARMITAGE, P., BRESLOW, N.E., COX, D.R., HOWARD, S.V., MANTEL, N., McPHERSON, K., PETO, J. & SMITH, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation on each patient. I. Introduction and design. *Br. J. Cancer*, **34**, 585–612.

POCOCK, S.J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics*, **35**, 183–197

SIMON, R. (1979). Restricted randomization designs in clinical trials. *Biometrics.*, **35**, 503–512.

SIMON, R. (1980). Patient heterogeneity in clinical trials. *Cancer Treat. Rep.*, **64**, 405–410.

SIMON, R. (1982). The design and conduct of clinical trials in oncology. In *Principles and Practice of Oncology*. eds. DeVita, V.T. Jr., Hellman, S. & Rosenberg, S.A., pp 198–225. Philadelphia: J.B. Lippincott.

TUKEY, J.W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science*, **198**, 679–684.

WHITE, S.J. & FREEDMAN, L.S. (1978). Allocation of patients to treatment groups in a controlled clinical study. *Br. J. Cancer*, **37**, 849–857.